

Реклассификация (Reclassification)

Синонимы: Переклассификация, Повторная классификация

Задача классификации решается с помощью обучения с учителем. Это означает, что в обучающем наборе данных для каждого наблюдения (примера) задана метка класса. При этом реализуется некоторое разбиение набора на классы, которое должен «научиться» воспроизводить классификатор в результате обучения.

Присвоение меток может осуществляться:

- **по факту** — когда метки определены самим условием задачи (например, заемщик допустил или не допустил просрочку).
- **с помощью эксперта** — когда метки назначает специалист предметной области на основе опыта решений аналогичных задач (например, какое количество покупок делает клиента лояльным или нелояльным).

В обоих случаях в задаче присутствует некоторый механизм назначения меток, называемый «учителем». В большинстве случаев он неидеален и может допускать ошибки. Однако слово «ошибка» здесь не совсем уместно, поскольку истинное (наилучшее) разбиение на классы, как правило, вообще неизвестно.

В результате разметка, сформированная учителем, оказывается смещенной относительно реальной. Классификатор, обученный на таких данных, может выдавать результаты, которые формально корректны с точки зрения алгоритма обучения, но далеки от действительности. В итоге это может привести к принятию неудачных управленческих решений.

Таким образом, при формировании обучающих данных для классификатора возникает два разбиения на классы: реальное (как правило, неизвестное) и созданное учителем, на основе которого проводится обучение. Поэтому в машинном обучении актуальна задача модификации разметки обучающего набора с целью уменьшения несоответствия между этими разбиениями. Эта задача получила название **реклассификация**.

Фактически реклассификация заключается в переназначении ранее заданных меток класса для улучшения работы классификатора. При этом «улучшение» понимается не как повышение точности, а как изменение воспроизводимой им структуры классов.

Необходимость в реклассификации может возникнуть в двух случаях:

- когда результаты работы классификатора изначально не соответствуют ожиданиям, хотя все оценки его точности высокие. Это означает, что структура классов в разметке обучающих данных изначально смещена относительно фактической.

- когда классификатор на первых этапах эксплуатации работает хорошо, но со временем его качество ухудшается. Это происходит из-за смещения фактической структуры классов относительно той, на которой он обучался.

На практике задача реклассификации сводится к поиску нового разбиения обучающего набора данных на классы, которое минимизирует некоторую целевую функцию. При этом могут появляться новые классы, исчезать существующие, а также происходить разделение или объединение ранее заданных классов.

В качестве целевой функции, например, можно использовать число наблюдений, изменивших метку класса при очередном разбиении. Поиск новых разбиений завершается, когда это число становится минимальным.